Apache Hive

Sekilas mengenai Apache Hive

Kalamangga.Net

Bagian Riset dan Pengembangan

Februari 2016

Outline

Batasan

- Perangkat
  - Distribusi Cloudera : CDH 5.2.4
  - OS : CentOS 6.5
- Penempatan data pada HDFS
- Web
  - https://hive.apache.org
  - https://cwiki.apache.org/confluence/display/Hive

Outline

Apache itu Hive?
Apache itu Hive?

- Sebuah proyek pada Yayasan Apache yang memiliki fungsi sebagai gudang data *data warehouse*.
- Dikembangkan di atas platform Apache Hadoop.
- Pada awal pengembangan merupakan sub-proyek dari Apache Hadoop.

Outline

Bagaimana penggunaan Hive?
Bagaimana penggunaan Hive?

- Umumnya digunakan pada proses ETL.
- Memberikan struktur data pada berbagai format berkas.
- Mengakses data langsung dari HDFS atau HBase.
- Sebagai gudang data.
  - Manajemen data.
  - Analisis data.

Outline

Bagaimana data dikelola?

- Data dapat dikelompokkan ke dalam *database*.
- Bila tidak ditentukan, *database* '`default`' akan digunakan.
- Data disimpan pada HDFS di lokasi sesuai konfigurasi '`hive.metastore.warehouse.dir`', konfigurasi standar menunjuk lokasi '`/user/hive/warehouse`'

Bagaimana data dikelola?

- Tabel pada Hive mirip dengan konsep tabel pada RDBMS.
- Setiap tabel diasosiasikan dengan sebuah direktori pada HDFS.
  - Misal : tabel '`pegawai`' pada *database* `default` diasosiasikan dengan direktori '`/user/hive/warehouse/pegawai`' di HDFS.

Bagaimana data dikelola?

- Tabel internal.
  - Tabel yang dimiliki oleh Hive.
  - Pengaturan tabel dan data sepenuhnya oleh Hive.
- Tabel eksternal.
  - Hive hanya dapat mengatur tabel.
  - Data diatur oleh perangkat / mekanisme lain.

Outline

Apa saja tipe data yang didukung?
Apa saja tipe data yang didukung?

- Numerik.
  - TINYINT.
  - SMALLINT.
  - INT.
  - BIGINT.
  - FLOAT.
  - DOUBLE.
  - DECIMAL.

Apa saja tipe data yang didukung?
Apa saja tipe data yang didukung?

- Waktu.
  - TIMESTAMP.
  - DATE.
- String.
  - STRING.
  - VARCHAR.
  - CHAR.
- Boolean.

Apa saja tipe data yang didukung?
Apa saja tipe data yang didukung?

- Binary.
- Kompleks.
  - Arrays.
  - Maps.
  - Structs.
  - Named Struct.
  - Union.
- Boolean.

Pengenalan
Pengenalan

- Apache Hive dapat melakukan *query* pada data yang disimpan dalam HDFS.
- Untuk melakukan *query*, Hive menyediakan bahasa SQL yang disebut `Hive Query Language` disingkat (HiveQL atau HQL).

Outline

DDL
Data Definition Language

- Database.
  - CREATE DATABASE.
    - COMMENT
    - LOCATION
    - WITH DBPROPERTIES
  - SHOW DATABASE.
    - LIKE '*'
  - DESCRIBE DATABASE.
  - USE.
  - DROP.
  - ALTER
    - SET DBPROPERTIES
    - SET OWNER USER

DDL
Data Definition Language

- Tabel.
  - CREATE TABLE.
    - ROW FORMAT DELIMITED
    - FIELDS TERMINATED BY
    - COLLECTION ITEMS TERMINATED BY
    - MAP KEYS TERMINATED BY
    - STORED AS
    - AS (CTAS)
    - LIKE
    - PARTITIONED BY
  - INSERT INTO TABLE
  - DROP TABLE

Outline

DSL
Data Selection Language

- SELECT.
  - FROM
  - WHERE
  - LIMIT
- Sub Query
  - IN
  - NOT IN
  - EXIST

DSL
Data Selection Language

- Join.
  - INNER JOIN
  - LEFT OUTER JOIN
  - RIGHT OUTER JOIN
  - FULL OUTER JOIN
  - CROSS JOIN
- UNION ALL.

Outline

DML
Data Manipulation Language

- LOAD DATA.
  - LOCAL
  - INPATH
  - OVERWRITE
  - INTO
- INSERT.
  - INTO TABLE
  - OVERWRITE
  - LOCAL DIRECTORY

DML
Data Manipulation Language

- EXPORT TABLE ... TO ...
- IMPORT TABLE ... FROM ...
- ORDER.
- SORT.
- Built-in Functions.
  - SHOW FUNCTIONS
  - DESCRIBE FUNCTIONS
    - EXTENDED

Outline

Data Aggregation

- COUNT.
- MAX.
- MIN.
- AVG.
- GROUPING SETS.
- ROLLUP.
- CUBE.
- GROUP BY.

Data Aggregation

- Kondisional.
  - CASE WHEN
  - COALESCE
  - IF

Data Aggregation

- Fungsi analisis.
  - RANK
  - DENSE_RANK
  - ROW_NUMBER
  - CUME_DIST
  - PERCENT_RANK

Data Aggregation

- Fungsi analisis.
  - NTILE
  - LEAD
  - LAG
  - FIRST_VALUE
  - LAST_VALUE

Outline

UDF
User-Defined Function

- Temporary.
  - CREATE TEMPORARY FUNCTION function_name AS class_name
  - DROP TEMPORARY FUNCTION [IF EXISTS] function_name
- Permanen.
  - CREATE FUNCTION [db_name].function_name AS class_name [USING JAR|FILE|ARCHIVE 'file_uri' [,USING JAR|FILE|ARCHIVE 'file_uri']]
  - DROP FUNCTION [IF EXISTS] function_name
- Reload.
  - RELOAD FUNCTION

UDF
User-Defined Function

- UDF (*User Defined Function*)
  - Berjalan pada tiap baris dan menghasilkan keluaran untuk tiap barisnya.
- UDAF (*User Defined Aggregate Function*)
  - Berjalan pada tiap baris atau kelompok baris dan menghasilkan keluaran untuk tiap baris atau kelompok baris yang didefinisikan.
- UDTF (*User Defined Table-Generating Function*)
  - Berjalan pada tiap baris atau kelompok baris dan menghasilkan keluaran berupa tabel.

UDF
User-Defined Function

- Langkah pembuatan
  - Buat fungsi dalam bahasa Java
  - `Compile` dan `pack` menjadi **JAR**
  - Load **JAR** ke **HDFS**
  - Daftarkan **JAR**
    - `ADD JAR /path/di/hdfs/nama_file.jar`
  - Definisikan fungsi
    - `CREATE TEMPORARY FUNCTION nama_fungsi AS net.kalamangga.dev.nama_fungsi`
  - Gunakan dalam *query*
    - `SELECT nama_fungsi(kolom1) FROM nama_tabel`

EoF
End of File

Terima Kasih

Disiapkan oleh :
Yudha H Tejaningrat
yht@kalamangga.web.id